Transportation Research Part C 62 (2016) 21–34

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# A spatiotemporal correlative *k*-nearest neighbor model for short-term traffic multistep forecasting

Pinlong Cai [a,b], Yunpeng Wang [a,b], Guangquan Lu [a,b,*], Peng Chen [a,b], Chuan Ding [a,b], Jianping Sun [c]

[a] Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, SiPaiLou #2, Nanjing 210096, China
[c] Beijing Transportation Research Center, Beijing 100073, China

ARTICLE INFO

*Article history:*
Received 8 May 2015
Received in revised form 27 September 2015
Accepted 7 November 2015
Available online 7 December 2015

*Keywords:*
Short-term traffic forecasting
*k*-nearest neighbor model
Spatiotemporal correlation
Gaussian weighted Euclidean distance

ABSTRACT

The *k*-nearest neighbor (KNN) model is an effective statistical model applied in short-term traffic forecasting that can provide reliable data to guide travelers. This study proposes an improved KNN model to enhance forecasting accuracy based on spatiotemporal correlation and to achieve multistep forecasting. The physical distances among road segments are replaced with equivalent distances, which are defined by the static and dynamic data collected from real road networks. The traffic state of a road segment is described by a spatiotemporal state matrix instead of only a time series as in the original KNN model. The nearest neighbors are selected according to the Gaussian weighted Euclidean distance, which adjusts the influences of time and space factors on spatiotemporal state matrices. The forecasting accuracies of the improved KNN and of four other models are compared, and experimental results indicate that the improved KNN model is more appropriate for short-term traffic multistep forecasting than the other models are. This study also discusses the application of the improved KNN model in a time-varying traffic state.


## 1. Introduction

Real-time traffic data can be obtained with the development of intelligent traffic systems (García-Ortiz et al., 1995). However, lag time is detected when such data are applied to formulate strategies for managing and controlling traffic (Smith et al., 2002). Therefore, short-term traffic conditions must be forecasted according to available real-time data. Short-term traffic forecasting has become an interesting research topic in this field. Administrators can manage traffic networks and effectively ensure normal operation with the aid of reliable forecasting data. Travelers can also decide on departure time or travel routes easily.

Several statistical models have been applied extensively in short-term traffic forecasting, including the time series model (Williams and Hoel, 2003; Lee and Fambro, 2007), Kalman filter model (Guo et al., 2014), nonparametric regression method (Smith and Demetsky, 1997; Smith et al., 2002; Zheng and Su, 2014), Bayesian model (Wang et al., 2014), support vector machine regression model (Zhang and Chen, 2010; Hu et al., 2011; Wang and Shi, 2013), and hidden Markov model (Qi and Ishak, 2014). Many artificial intelligence algorithms have also been used effectively to this end. The most typical model established is the neural network model (Smith and Demetsky, 1994; Xie and Zhang, 2006; Dong et al., 2010; Hou, 2011;

* Corresponding author at: Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China.

http://dx.doi.org/10.1016/j.trc.2015.11.002

Abdi et al., 2012; Leng et al., 2013; Ma et al., 2015a). Besides, Huang and Sadek (2009) proposed a spinning network model that is similar to human memory. Zhang et al. (2014a) developed a hierarchical fuzzy rule-based system that is optimized by genetic algorithms.

Considerable research has concentrated on comparing an alternative forecasting model with other models given the ready availability of different types of data and the unique features of various models (Vlahogianni et al., 2014). Artificial intelligence algorithms can overcome several problems (e.g., data failure) and provide black box solutions. Model performance constrains the quality of trained data, and these models are difficult to extend from one application to another (Van and Van, 2012). By contrast, the traditional statistical models are simple to implement and are applicable to numerous road segments; however, these models hardly forecast accurately when used alone. The nonparametric regression method has more portability, higher accuracy, and a simpler structure than the parametric models do (Smith and Demetsky, 1997; Zheng and Su, 2014); as a data-driven approach, the former is also suitable for short-term traffic forecasting in urban road networks because this method adapts to the complexity of traffic signals through flexible restructuring (Vlahogianni et al., 2014). An example of the characteristic nonparametric regression method is the $k$-nearest neighbor (KNN) model, which is easy to implement because the process of training data and estimating parameters is simple. Nonetheless, the search algorithm and method of forecasting result integration in this model should be improved.

The majority of previous studies conducted single-step forecasting depending on the limited data regarding a single road. The duration of such forecasting is less than 15 min, which is generally relatively short to help travelers complete one trip on the currently complex road networks. Several scholars have recently attempted to investigate multistep forecasting; however, the performance of such forecasting deteriorates rapidly with an increase in the number of steps when traditional forecasting methods are employed. Thus, researchers analyzed the relationship among road segments by considering much spatiotemporal data collected from several road segments in a road network. Min and Wynter (2011) considered the distance and average speed of the links in reflecting the spatial characteristics of a road network; these indicators remained accurate for up to 1 h in 12 time steps. Considering the influence of multiple links, Sun et al. (2012) proposed the Bayesian classical model based on the Gaussian regression process for short-term traffic forecasting in urban road networks. Zhang et al. (2014b) established a neural network model of radial basic function by analyzing the traffic flow relationship between a specific road segment and other road segments. Haworth and Cheng (2012) predicted travel time in the central London section by combining the nuclear regression model with the KNN model. Kamarianakis et al. (2012) enhanced the classic time series model and examined a spatiotemporal correlation by analyzing traffic flow variables and nonlinear dynamics. Zou et al. (2014) merged spatial and temporal travel time information to predict travel time within 1 h. The aforementioned researchers considered the spatiotemporal data of nearby road segments for short-term traffic forecasting; however, these scholars were unable to quantize the spatiotemporal correlation among road segments clearly in their forecasting models. Thus, an improved KNN model is proposed in the present study based on the spatiotemporal correlation of road segments.

The present study applies a new criterion of equivalent distances to redefine the contact among road segments and uses spatiotemporal state matrices to identify traffic states. Then, the proposed model enhances the computations of nearest neighbor distance and the integration of forecasting results through the Gaussian weighted method to overcome the defect of the original KNN model. Peak and off-peak times are set within one day given the dynamic flows in a road network. In addition, a time-varying model is reasonably used to forecast short-term traffic states involving different traffic characteristics in various periods. Finally, a deviation compensation method is introduced to adjust the forecasting result further. The maximum forecasting time in the study is 1 h (12 time steps).

This paper is organized as follows: Section 2 proposes the improved KNN model based on spatiotemporal correlations. The equivalent distance, spatiotemporal state matrix, and Gaussian weighted methods are also introduced in this section. Section 3 determines the parameters in the model and compares the performance of the improved KNN model with that of other models with the real data collected. Section 4 discusses the application of the improved KNN model in a time-varying traffic state and the deviation adjustment of the forecasting results. Section 5 concludes the study and proposes directions for future research.

## 2. Methodology

### 2.1. Original KNN model

The KNN model for short-term traffic forecasting aims to identify the current state of the traffic network and integrates generations of similar historical states as forecasting results. The specific steps of the original KNN model are described as follows.

First, historical data are used to build the sample database. Either traffic flow or vehicle speed is usually selected as the critical parameter. An appropriate vector space is also defined to describe the current and historical traffic states. Then, the Euclidean distances between all sample and current data are calculated to generate the $k$-sample data, the distances among which are regarded as the KNNs. Finally, future traffic states are forecasted by averaging generations of KNNs (Smith and Demetsky, 1997).

The original KNN model has been applied in many studies, and several analogous weaknesses should be improved upon. The results of the original KNN model are usually hysteretic in time series and lack prediction accuracy because this model